EGC 2016
REIMS

16ème Conférence Internationale
Francophone sur l'Extraction
et la Gestion des Connaissances

18 - 22 Janvier 2016
http://egc2016.univ-reims.fr

Actes de l'atelier GAST – **Gestion et Analyse de données Spatiales et Temporelles**

Éric Kergosien (GERiiCO, Université Lille 3)
Thomas Guyet (IRISA-Inria/AGROCAMPUS-OUEST)
Christian Sallaberry (LIUPPA, Université de Pau et des Pays de l'Adour)

http://gt-gast.irisa.fr/gast-2016/

Mardi 19 janvier 2016 Reims

PRÉFACE

La gestion et l'analyse de données spatiales connaît une dynamique forte grâce au développement de l'estampillage spatial ou temporel des données. Pour répondre aux besoins d'exploration approfondie des données et d'exploitation des informations qu'elles contiennent, des méthodes et outils spécifiques sont requis.

Les objectifs des ateliers GAST (Gestion et Analyse des données Spatiales et Temporelles) concernent notamment la prise en compte de la quantité et de la richesse des données spatiales et/ou temporelles diffusées dans les contenus numériques. La prise en considération de la variété des données numériques (sources, contenus, types de documents, etc.) est également une véritable problématique mais c'est aussi une force dans la quête d'identification de la connaissance. Autrement dit, comment identifier, extraire, structurer et mettre à disposition des acteurs (experts, usagers lambda, etc.) des connaissances s'appuyant sur des données spatiales et temporelles à partir des contenus numériques hétérogènes disponibles ? Ces différents défis lèvent des verrous scientifiques multidisciplinaires qui sont traités au sein la série d'ateliers GAST.

Ces actes regroupent les soumissions acceptées à l'atelier GAST en 2016 dans le cadre de la conférence Extraction et Gestion des connaissances (EGC) organisée à Reims par C. de Runz. Cet atelier est un rendez-vous annuel fédérateur, convivial et scientifique riche de l'ensemble de la communauté s'intéressant à la gestion et à l'analyse de données spatiales et temporelles. Cette année, l'atelier a été organisé en trois temps : une présentation invitée de Danielle Ziebelin sur les données spatio-temporelles ouvertes et liées, puis un ensemble de présentations orales des articles retenus pour l'atelier et finalement un temps dédié à la discussion avec l'ensemble des participants. Nous espérons que le lecteur qui n'a pu y assister trouvera toutes les informations dans les articles de ce volume.

L'article "Identification automatique des types de relations spatiales dans les textes", de Sarah Zenasni, Eric Kergosien, Mathieu Roche et Maguelonne Teisseire apporte un éclairage sur la découverte de connaissances à partir de documents textuels et, plus particulièrement, l'identification d'informations spatiales. La méthode proposée combine des approches de fouille de textes pour identifier les types de relations spatiales de façon automatique. Les résultats des expérimentations réalisées sur un corpus en anglais sont également présentés et discutés.

L'article "Approche pour l'élaboration d'un modèle chronotopique urbain", de Alain Guez et Francis Rousseaux, s'inscrit dans le cadre de l'étude chronotopyque d'un territoire. Ce type d'étude vise à analyser, avec un point de vue géographique, les rythmes de présence et co-présence des résidents – et des habitants temporaires – en fonction des activités, des horaires et l'organisation de la ville. La démarche entreprise dans le cadre de ce travail vise à mobiliser de l'information disponible (ici des horaires d'ouvertures/fermetures d'activité collectée à partir de différences sources)

pour répondre à ces questions de géographie. Il partage ainsi des questionnements sur les représentations conjointes des dimensions spatiales et temporelles. Des premières propositions de représentations de cartes et de caractéristiques temporelles sont faites, mais l'article ouvre surtout sur les défis que représentent leurs questions en terme d'exploitation de données (massives) et de leur analyse automatique.

L'article "Cognisearch Business : un service de recherche d'information d'entreprises sur le web", d'Armel Fotsoh, Annig Le Parc-Lacayrelle et Tanguy Moal, vise la construction de "cartes d'identités" d'entreprises à partir de données du Web. Le modèle d'entité entreprise décrit des informations d'immatriculation (SIREN), des coordonnées (Web, téléphone, adresse) et un contexte (métier, activité, produit). La chaine de traitement comprend des modules de filtrage de sites Web, d'annotation des informations spatiales et thématiques, d'enrichissement des ressources externes, d'indexation et recherche d'information. Un prototype met en œuvre ces propositions.

L'article "La confiance est dans l'air ! Application à l'identification des parcours hospitaliers", de Yves Mercadier, Jessica Pinaire, Jérôme Azé, Sandra Bringay et Maguelonne Teisseire, vise l'identification de séquences fréquentes d'événements ordonnés. Les verrous concernent notamment l'extraction de motifs séquentiels et la notion de confiance appliquée à ces motifs. Le domaine d'application est la prédiction de la trajectoire de patients ayant eu un infarctus du myocarde entre 2009 et 2013. Les résultats obtenus sont discutés par un spécialiste.

L'article "Analyse multi-échelles de référentiels vectoriels via SOLAP", de Marie-Dominique Van Damme et Sébastien Mustière, concerne l'exploration de larges volumes de données hétérogènes. L'approche expérimentale propose d'intégrer une base de données topographique dans un SOLAP. Elle a pour objectif d'analyser et agréger, via des requêtes ad-hoc, les données à grande échelle et vectorielles de l'IGN. Ce travail permet déjà à travers des zooms sur les dimensions spatiales et des sélections temporelles de découvrir différents phénomènes. Ces premiers résultats, via les outils SOLAP, contribuent à la modélisation multidimensionnelle et à l'étude de données vecteurs de l'IGN à des échelles variées.

Ces articles montrent une large étendue des recherches actuelles dans le domaine de la gestion et de l'analyse de l'information spatiale et temporelle. Nous y trouvons avec plaisir des thématiques aussi différentes que le traitement automatique des langues et la fouille portant sur des données temporelles, spatiales et thématiques. Tout ceci correspond aux intérêts premiers de GAST. Au travers de cet atelier, nous espérons que les orateurs, les auditeurs et les lecteurs constatent la complexité que continue de poser l'information temporelle et spatiale, qu'ils voient les défis qui se posent encore aux chercheurs.

Nous tenons à remercier tous les auteurs pour leurs propositions d'articles ainsi que les membres du comité de lecture qui ont su respecter les contraintes imposées par le

planning serré d'un atelier et dont les relectures ont été de qualité pour l'ensemble des articles. Nous remercions également chaleureusement Danielle Ziebelin, professeure à l'Université de Joseph Fourier, pour son intervention en tant que conférencière invitée à la journée GAST'2016. En espérant que ces articles vous apporteront de nouvelles perspectives sur la gestion et l'analyse de données spatiales et temporelles, nous vous souhaitons une bonne lecture.

Eric Kergosien
Université Lille-3/GERiiCO

Thomas Guyet
Agrocampus-Ouest/IRISA-Inria

Christian Sallaberry
Université de Pau/LIUPPA

Membres du comité de lecture

Peggy Cellier - IRISA, Rennes
Christophe Claramunt - Ecole Navale, Brest
Géraldine Del Mondo - INSA, Rouen
Thomas Devogele - LI, Tours
Catherine Domingues - IGN, Saint-Mandé
Frédéric Flouvat - PPME, Nouméa
Thierry Joliveau - CRENAM, EVS, Saint-Etienne
Éric Kergosien - GERIICO, Lille
Florence Le Ber - ENGEES, Strasbourg
Simon Malinowski - IRISA, Rennes
Thomas Guyet - AGROCAMPUS-OUEST/IRISA, Rennes
Simon Malinowski - IRISA, Rennes
Nicolas Meger - LISTIC, Annecy
René Quiniou - Inria, Rennes
Sébastien Mustière - IGN, Saint-Mandé
Mathieu Roche - CIRAD, Montpellier
Fatiha Saïs - LRI, Paris
Nazha Selmaoui-Folcher - PPME, Nouméa
Christian Sallaberry - LIUPPA, Pau
Nazha Selmaoui - PPME, Nouméa
Maguelonne Teisseire - IRSTEA, Montpellier
Karine Zeitouni - PRISM, Versailles

TABLE DES MATIÈRES

Présentation invitée

Articles de l'atelier

Index des auteurs

Identification automatique des types de relations spatiales dans les textes

Sarah Zenasni[*,***] Eric Kergosien [**]
Mathieu Roche [*,***] Maguelonne Teisseire [*,***]

[*]UMR Tetis (IRSTEA, CIRAD, AgroParisTech), France
prénom.nom@teledetection.fr
[**]GERiiCO, Univ. Lille 3, France
prénom.nom@univ-lille3.fr
[***]LIRMM, CNRS, Univ. Montpellier, France
prénom.nom@lirmm.fr

Résumé. La découverte de connaissances à partir de documents textuels, en particulier l'identification d'informations spatiales, est une tâche difficile due à la complexité de l'analyse des textes écrits en langage naturel. Dans nos travaux, nous proposons une méthode combinant des approches de fouille de textes pour identifier les types de relations spatiales de façon automatique. Les résultats des expérimentations réalisées sur un corpus en anglais sont présentés et discutés.

1 Introduction

L'extraction d'information spatiale prend une importance croissante non seulement sur les entités spatiales (ES), mais aussi sur les relations entre entités spatiales. Ces dernières se sont avérées complexes à saisir, à définir et donc à modéliser. Le travail présenté dans cet article se situe dans un tel contexte, l'objectif est de découvrir le type des relations entre les entités spatiales exprimées dans les textes. Nous nous concentrons plus particulièrement sur trois types de relations spatiales : région (par exemple, "leading up"), direction (par exemple, "going up") et distance (par exemple, "near"). La suite de cet article est organisée de la façon suivante. La section 2 présente une brève introduction des travaux existants en extraction d'information spatiale. Puis, nous décrivons, en section 3, les deux approches proposées et leur combinaison. Nous détaillons, en section 4, le protocole expérimental et les résultats obtenus. Finalement la section 5 présente la conclusion et les perspectives de nos travaux.

2 État de l'art

De nombreux travaux s'intéressent à l'identification d'Entités Nommées (EN), et plus particulièrement d'Entités Spatiales à partir de données textuelles (Nadeau et Sekine, 2007). Ces approches s'appuient sur des méthodes linguistiques (par patrons d'extraction par exemple) (Maurel et al., 2011) et / ou sur des méthodes statistiques (Velardi et al., 2001). Ces techniques sont intéressantes pour l'identification d'Entités Spatiales, mais elles ne permettent pas

d'identifier l'information spatiale de manière plus exhaustive. Une meilleure représentation de la connaissance spatiale peut être obtenue en considérant les informations sur les relations spatiales. Globalement, les relations peuvent être identifiées par des calculs de similarité entre des contextes syntaxiques (Grefenstette, 1994), par prédiction à l'aide de réseaux bayésiens (Weissenbacher et Nazarenko, 2007), par des techniques de fouille de textes (Grčar et al., 2009). Cependant, ces approches ne permettent pas toujours d'identifier la sémantique de la relation. Nos travaux s'inscrivent dans ce contexte et visent à reconnaitre de façon automatique le type de la relation spatiale étudiée. Plus précisément, nous proposons une méthode hybride, combinant des informations lexicales et contextuelles et une approche de fouille de textes pour prédire le type de relations spatiales.

3 Prédiction du type de relation spatiale

Dans la suite de cet article, nous nous appuierons sur les deux phrases ci-dessous pour lesquelles nous cherchons à prédire la classe des relations spatiales (en gras).
Phrase 1 : Stairs are **leading up** to the entrance.
Phrase 2 : Four locals are **sitting on** a bench in a canteen kitchen , **leaning on** a red brick wall.

3.1 Par comparaison de chaînes de caractères

Parmi les nombreuses mesures de similarité existantes, nous avons choisi deux méthodes ***String Matching (SM)*** (Maedche et Staab, 2002) et ***Lin*** (Lin, 1998) qui sont classiquement utilisées dans la littérature car elles produisent des résultats pertinents (Duchateau et al., 2008).

3.1.1 String Matching

SM est une mesure lexicale fondée sur la ***distance de Levenshtein*** (notée L) (Navarro, 2001), elle calcule la somme minimale du coût des opérations **suppression, insertion, remplacement** nécessaires pour transformer une chaîne de caractères *Ch1* en *Ch2*. À partir de

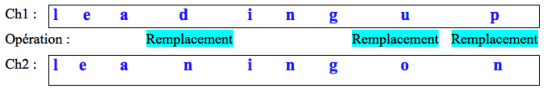

FIG. 1 – *Distance de Levenshtein pour les relations "leading up" et " leaning on".*

l'exemple présenté dans la Figure 1, nous obtenons L(leading up, leaning on)=3. En effet, il y a trois opérations permettant de passer de la chaîne "leading up" à "leaning on". Après avoir calculé la distance L, nous appliquons la formule (1) pour calculer la valeur de SM, normalisée entre 0 et 1.

$$SM(Ch1, Ch2) = max[0; (min(|Ch1|, |Ch2|) - E(Ch1, Ch2))/min(|Ch1|, |Ch2|)] \quad (1)$$

À partir de l'exemple des phrases 1 et 2, SM(leading up, leaning on)=$max[0, (10-3)/10] = 0.70$. Sur la base de ces mesures, nous avons retourné pour chaque relation candidate pour

laquelle nous voulons prédire la classe, les similarités obtenues avec l'ensemble des relations de l'ensemble d'apprentissage (voir l'explication détaillée en section 4). Nous déterminons ainsi les K relations les plus proches afin de prédire la classe à associer à la relation candidate (algorithme des K plus proches voisins $KPPV$). Pour les cas particuliers tels que les relations composées de deux mots dont le deuxième mot est une relation spatiale **next to, standing in...**, nous faisons l'hypothèse que ces relations sont du même type que celui des relations **to, in...**

3.1.2 Lin

La mesure Lin est une mesure de similarité fondée sur l'identification des n-grammes de caractères. Généralement, la valeur de n varie entre 2 et 5. En posant $n = 3$ (tri-grammes notée tr), nous obtenons le résultat ci-dessous en reprenant l'exemple de la section 3 :
tr (leading up) = {**lea**,ead,adi,din,**ing,ng** ,g u, up} = 8
tr (leaning on) = {**lea**,ean,ani,nin,**ing,ng** ,g o, on} = 8
tr (leading up) ∩ **tr** (leaning on) = 3
La formule 2 présente la mesure *Lin* normalisée entre 0 et 1 :

$$Lin(Ch1, Ch2) = \frac{1}{[1 + |tr(Ch1)| + |tr(Ch2)| - 2 \times |tr(Ch1) \cap tr(Ch2)|]} \quad (2)$$

À partir de l'exemple des phrases 1 et 2, Lin(leading up, leaning on) = $\frac{1}{[(1+8+8)-(2\times3)]} = 0.09$.
Sur la base de cette mesure de similarité, nous avons également appliqué l'algorithme $KPPV$ qui retourne la classe majoritaire pour la relation candidate. Notons que les informations lexicales ne sont pas toujours suffisantes. En effet, deux expressions peuvent être lexicalement éloignées mais sémantiquement très proches. Pour résoudre un tel problème, nous proposons, dans la section suivante, de prendre en compte le contexte des relations pour prédire leur classe.

3.2 Par proximité contextuelle

À cette étape, nous faisons l'hypothèse que les mots présents autour des relations (toute la phrase ou les n mots autour de la relation), que nous nommons "*monde lexical*", vont nous permettre d'améliorer l'identification du type des relations spatiales. Nous nous appuyons ensuite sur une approche sac de mots "SDM", nous comparons différents facteurs de pondération : nombre d'occurrences, TF-IDF (Salton et Buckley, 1988) et la confiance (Agrawal et al., 1993) afin de sélectionner celui qui nous permet de construire le monde lexical le plus à même d'identifier le type de relation spatiale pertinent. Sur la base des trois pondérations des mots du monde lexical, nous mesurons la proximité fondée sur le cosinus entre les mondes lexicaux propres aux relations candidates et aux relations de l'ensemble d'apprentissage. Une fois l'ensemble des mesures de proximité calculées, nous appliquons l'algorithme $KPPV$ et nous affectons chaque relation candidate à la classe identifiée comme la plus proche.

3.3 Combinaison

Observant que toutes ces approches prises séparément restent imparfaites, nous proposons une méthode combinant les deux méthodes précédentes (par comparaison de chaînes de caractères et par proximité contextuelle). Dans le cadre de nos expérimentations (voir section

4), nous obtenons la liste des relations prédites pour chaque approche. En analysant qualitativement les résultats obtenus, nous remarquons que l'approche par comparaison de chaînes de caractères donne généralement de meilleurs résultats. Cependant l'approche par proximité contextuelle donne des résultats sensiblement meilleurs lorsque les relations se composent de plus de 4 termes. Au regard de cette première analyse, nous faisons l'hypothèse que si les relations se composent de plus de n [1] termes, nous privilégions la proximité contextuelle Cos, sinon nous choisissons l'approche par comparaison de chaînes de caractères *SM*. La section 4 décrit les résultats de nos expérimentations menées sur un corpus en langue anglaise.

4 Expérimentations

Pour mener à bien ces expérimentations, nous avons choisi un corpus en langue anglaise SPRL (Spatial Role Labeling) (Parisa et al., 2012) qui représente un benchmark reconnu dans le domaine. Le corpus est composé de 1213 phrases annotées. Nous avons procédé à une série d'expérimentations dans lesquelles nous avons fait varier les paramètres susceptibles d'influencer les résultats des mesures de performance : K pour l'algorithme de $KPPV$ et n paramètre de fenêtrage.

4.1 Évaluation de la proximité lexicale

Afin d'estimer l'efficacité des différentes méthodes, nous appliquons un processus de validation croisée. Dans notre cas, le corpus est divisé en 3 partitions et chaque partition contient 31 relations (18 régions, 10 directions, 3 distances). Le jeu d'apprentissage est constitué successivement de 2 des 3 partitions et le jeu de test permettant d'obtenir les résultats présentés est constitué de la partition restante. Le tableau 1 représente dans la colonne *String Matching 1* les résultats obtenus en terme d'exactitude (accuracy) à partir de la 1ère série d'expérimentations, en appliquant l'algorithme SM uniquement. La colonne *String Matching 2* représente les résultats obtenus en appliquant la règle présentée en section 3.1.1 de relations composées de deux mots. La mesure de similarité SM donne des résultats satisfaisants comparativement à la mesure Lin quelque soit la valeur de K avec un score de 0.82 d'exactitude.

K	*String Matching 1*	*String Matching 2*	*Lin*
1	0.77	**0.82**	0.75
3	0.74	0.79	0.73
5	0.73	0.76	0.69

TAB. 1 – *Résultats des mesures SM et Lin en terme d'exactitude.*

4.2 Évaluation de la proximité contextuelle

Dans cette série d'expérimentations, la prédiction des relations spatiales est effectuée sur la base de l'approche sac de mots classique avec suppression des "mots vides" [2]. Nous appliquons les deux contextes (toute la phrase, n mots autour de la relation) et nous évaluons l'approche

1. Nos expérimentations ont montré que $n = 4$ donne les résultats les plus pertinents
2. http ://xpo6.com/list-of-english-stop-words/

pour chaque monde lexical avec K' variant de 1 à 5. Dans le tableau 2, nous pouvons constater que le contexte n *mots autour de la relation* donne des résultats supérieurs à ceux du contexte *toute la phrase*. Le monde lexical fondé sur le TF-IDF avec $K' \geq 3$ donne des résultats satisfaisants. Comme conclusion de cette série d'évaluations, le meilleur score (exactitude de 0.67) est obtenu avec $K' = 5$ et $n = 2$.

K'		*toute la phrase*	*n termes autour de RS*		
			n = 1	n = 2	n = 3
1	nombre d'occurrences	0.61	0.62	0.62	0.60
	TF-IDF	0.56	0.60	0.51	0.53
	Confiance	0.56	0.62	0.62	0.60
3	nombre d'occurrences	0.62	0.60	0.63	0.58
	TF-IDF	0.45	0.63	0.66	0.63
	Conf	0.40	0.60	0.63	0.56
5	nombre d'occurrences	0.58	0.65	0.67	0.57
	TF-IDF	0.40	0.64	**0.67**	0.66
	Confiance	0.41	0.64	0.67	0.56

TAB. 2 – *Résultats de la méthode de proximité contextuelle (notée Cos) en terme d'exactitude.*

4.3 Combinaison

Dans cette section, nous présentons les résultats de la combinaison. Nous avons réalisé une série d'expérimentations pour identifier la combinaison de paramètres les plus adaptés, i.e. $K = 1$ pour SM et $K' = 5, n = 2$ pour Cos utilisant le monde lexical fondé sur TF-IDF. Ceci nous permet d'obtenir un score d'exactitude de 0.84. Ainsi, la combinaison des deux méthodes (comparaison de chaînes de caractères et prise en compte du monde lexical) se comporte mieux que chaque méthode individuellement.

5 Conclusion

Dans cet article nous avons proposé une analyse comparative de deux approches et de leur combinaison pour l'identification automatique du type des relations spatiales. Nous avons défini un monde lexical pour améliorer la prédiction. Puis nous avons proposé une méthode combinant plusieurs mesures de similarité et pondérations. Nos résultats montrent que la combinaison améliore la qualité de la prédiction. Cela nous permet d'explorer de nouveaux modes d'hybridation afin de tirer le meilleur parti des différentes approches (lexicales et contextuelles). Comme perspective, dans un premier temps, nous voulons étudier la généricité de la méthode. Pour cette raison nous voulons exploiter un corpus contenant 7000 documents de Midi Libre [3]. Dans un deuxième temps, nous voulons étudier l'adaptation de l'approche selon le type de textes traités (presse vs. réseaux sociaux/SMS). Ainsi, nous envisageons d'exploiter un corpus contenant plus de 88.000 SMS [4]. En effet, un tel corpus contient des expressions de

3. http ://www.lirmm.fr/ mroche/ANIMITEX/participants.html
4. http ://88milsms.huma-num.fr

spatialité spécifiques à la communication média et aux réseaux sociaux (par exemple : "je v a montpel"). Finalement, nous souhaitons également nous intéresser à la prédiction de la classe propre aux relations spatiales entre différents types d'entités nommées (personne, organisation, etc).

Références

Agrawal, R., T. Imielinski, et A. N. Swami (1993). Mining association rules between sets of items in large databases. In *Proc. of Int. Conf. on Manag. of Data (SIGMOD)*, pp. 207–216.

Duchateau, F., Z. Bellahsene, et M. Roche (2008). Improving quality and performance of schema matching in large scale. *Ingénierie des Systèmes d'Information 13*(5), 59–82.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Grčar, M., E. Klien, et B. Novak (2009). Using Term-Matching Algorithms for the Annotation of Geo-services. In *Knowl. Disc. Enh. with Sem. and Soc. Info.*, pp. 127–143.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. of the Fifteenth Int. Conf. on Machine Learning (ICML)*, pp. 296–304.

Maedche, A. et S. Staab (2002). Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Int. Conf. EKAW*, pp. 251–263.

Maurel, D., N. Friburger, J.-Y. Antoine, I. Eshkol-Taravella, et D. Nouvel (2011). Casen: a transducer cascade to recognize french named entities. *TAL 52*(1), 69–96.

Nadeau, D. et S. Sekine (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes 30*(1), 3–26.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Comp. Surv.*, 31–88.

Parisa, K., B. Steven, et M. Marie-Francine (2012). Semeval-2012 task 3: Spatial role labeling. pp. 365–373. ACL.

Salton, G. et C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Inf. Proc. Manage. 24*(5), 513–523.

Velardi, P., P. Fabriani, et M. Missikoff (2001). Using text processing techniques to automatically enrich a domain ontology. In *FOIS*, pp. 270–284.

Weissenbacher, D. et A. Nazarenko (2007). Identifier les pronoms anaphoriques et trouver leurs antécédents: l'intérêt de la classification bayésienne. In *Proc. of TALN*, pp. 145–155.

Summary

Knowledge discovery from texts, particularly the identification of spatial information is a difficult task due to the complexity of the texts written in natural language. In our work, we propose a method combining two statistical approaches (lexical and contextual analysis) and a text mining approach to identify the types of spatial relationships. Experiments conducted an english corpus are presented.